이 성 건 교수지도

석사학위청구논문

SAS/AF를 이용한 통계적 데이터 결합 시스템에 관한 연구

2008

성신여자대학교 대학원

통 계 학 과

김 희 라

SAS/AF를 이용한 통계적 데이터 결합 시스템에 관한 연구

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2007년 11월

성신여자대학교 대학원

통 계 학 과

김 희 라

인 준 서

김희라의 석사학위 논문으로 인준함.

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

논 문 개 요

최근 정보화 사회의 발전과 더불어 서로 다른 원천과 경로로부터 얻어지는 많은 양의 자료들 속에서 의미 있는 정보를 찾고자하는 방법들이 연구되어 왔다. 그 하나의 방법으로 원천이 서로 다른 데이터들을 통계적으로 결합하여 의미 있는 정보를 얻을 수 있는 “통계적 자료 결합 방법(statistical data matching)”이 제안되고 있다. 통계적 자료결합의 방법으로는 유클리드 거리 혹은 마할라노비스 거리를 이용하는 ‘최근접이웃(nearest neighbor)’방법과, 회귀분석을 이용하는 방법, 성향점수(propensity score)를 이용하는 방법 등이 있다. 이러한 통계적 자료 결합 방법은 결합하고자 하는 자료의 특성과 상황에 따라 적절히 사용되어야 한다.

본 연구에서는 기존에 제안되어 있는 통계적 결합 방법론들을 전문지식이 없는 연구자들도 쉽게 사용할 수 있도록 통계적 결합 시스템을 구축하였다. 구축 방법은 SAS 소프트웨어에서 제공되는 기능을 이용하였다. SAS 제품 중 응용프로그램을 개발할 수 있도록 지원하는 SAS/AF (Application Frame) 와 SCL(Screen Control Language)를 이용하여 두 개의 파일을 통계적 방법으로 의미 있게 결합하는 통계적 결합 시스템 “SDMS (Statistical Data Matching System)”를 구축한다. 개발된 시스템으로 통계청의 2007년 서울 도시가계조사 자료에 적용하여 결합을 수행해 보고 그 결과를 비교한다.

<목 차>

제1장 서 론

현대사회를 정보의 홍수라고 할 만큼 우리가 접하는 정보들은 여러 경로로부터 무분별하게 얻어지고 그 양과 규모가 커지고 있다. 그러나 자료의 수집환경은 열악해지고, 그에 따라 자료의 질이 떨어지는 것이 현실이다. 이러한 현실에서 의미 있는 정보를 찾을 수 있는 신뢰할만한 자료를 얻기는 어려우며 많은 비용과 시간이 필요하게 된다. 따라서 주어진 서로 다른 자료들을 결합하여 의미 있는 정보를 얻을 수 있는 자료 결합이 해결방안으로 제안되고 있다. 이는 자료를 재조사 하는 방법보다 비용과 시간을 절약할 수 있으며, 많은 문항으로 인한 조사 응답자의 부담을 줄여줌으로써 무응답이나 부정확한 응답의 문제를 해결하여 더욱 신뢰성 있는 자료를 얻을 수 있을 것이다.

자료 결합은 주민등록번호 등과 같은 식별 변수를 사용하여 값이 정확히 일치하는 개체를 찾아 결합하는 정확 결합(exact matching)방법과, 주민등록번호와 같은 식별 변수가 없는 경우 결합하고자 하는 개체와 가장 유사한 개체를 찾아서 결합하는 통계적 결합(statistical matching)방법으로 나뉠 수 있다. 정확 결합 방법은 자료를 완벽하게 결합할 수 있다는 장점이 있으나, 대부분의 자료들은 개인정보 보호 등의 이유로 주민등록번호와 같은 개인정보를 대개 포함하지 않고 있다. 따라서 서로 다른 원천으로부터 얻어진 식별변수가 없는 자료들을 결합하기 위해서는 통계적 결합 방법이 주로 사용된다.

또한 통계적 결합은 자료의 결합에 사용되는 통계적 방법들에 따라 여러 가지 방법으로 나뉘어 그 배경을 이해하는 데에 통계적 지식이 필요하고, 자료를 결합하기 위해서 SAS 등의 통계 프로그램이 사용된다. 그러나 비전문가의 경우 통계적 지식이 부족하고 통계 프로그램을 잘 다루지 못하는 경

우가 많아 직접 자료를 결합 하는 데에 어려움이 있다. 따라서 누구나 손쉽게 자료를 결합할 수 있는 자료 결합 시스템이 필요하다.

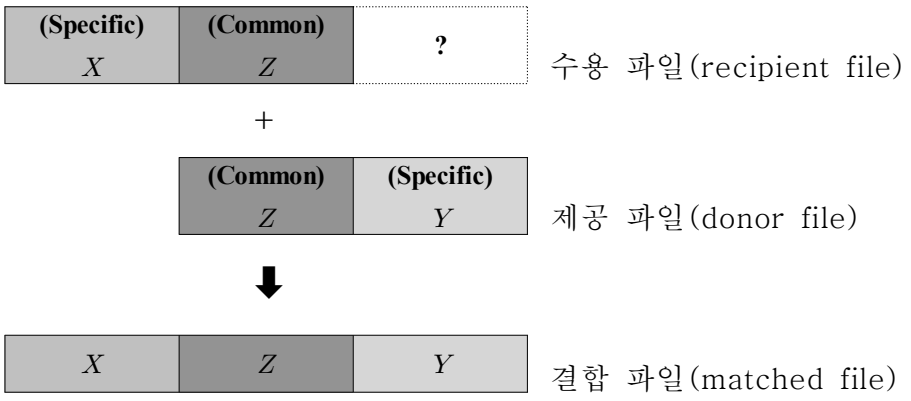
본 논문의 2장에서는 통계적 결합의 의미와 구조에 대해 자세히 알아보고 3장에서는 최근접이웃(nearest neighbor matching)방법이나 성향점수(propensiby score matching)방법 등 통계적 결합의 여러 가지 기법에 대해 살펴본다. 4장에서는 SAS/AF 와 SCL을 이용하여 누구나 손쉽게 자료를 결합할 수 있는 자료 결합 시스템을 구현하고, 그 시스템을 이용하여 실제 자료를 결합한 후 비교해 본다. 5장에서는 결론 및 향후 연구방향에 대하여 논의한다.

제2장 통계적 결합

2.1 통계적 결합의 의미

Saporta(2002)는 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측값이 존재할 경우 다른 원천으로부터 얻어지는 데이터와 정보(information)를 통합시키는 것이 통계적 결합이라고 정의하고 있다.

통계적 결합은 완전히 다른 양쪽 자료로부터 이루어진다. 결합에 의해 변수가 추가될 파일은 수용 파일(recipient file), 변수를 제공할 파일은 제공파일(donor file)이라 한다. 그리고 이 양쪽 데이터에서 공통으로 관측되는 변수를 공통변수(common variable)라 하고 각각의 파일에서 어느 한쪽에만 존재하는 변수를 고유변수(specific variable)라고 한다. 아래 그림과 같이 고유변수 X, Y 공통변수를 Z라고 표시 할 때 공통변수 Z를 이용하여 제공파일의 고유변수 Y를 수용파일에 추가하여 결합된 파일을 결합 파일(matched file)이라 한다.



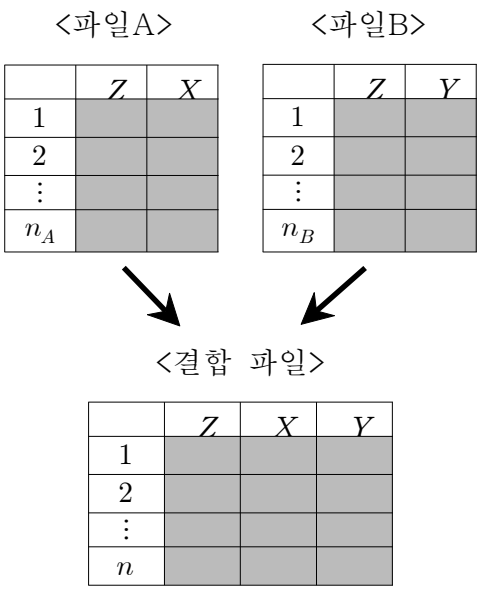
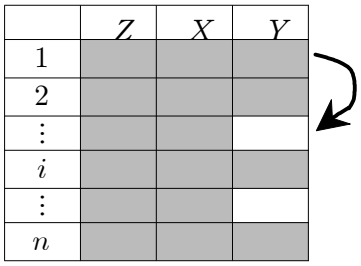
<그림 2-1> 통계적 결합(statistical matching)

2.2 결측치 대입과 통계적 결합의 비교

통계적 결합은 변수의 값을 기존에 존재하는 값으로 만들어주는 것이고, 결측값 대체는 데이터에 존재하지 않더라도 그 변수를 대표할 수 있는 값으로 채워주는 방법이다.

자료에 결측치가 있을 때 이를 처리하는 방법에는 여러 가지가 있지만, 관측된 변수를 이용하여 그와 비슷한 값을 갖는 개체의 변수 값으로 결측값을 대입하는 방법은 통계적 결합 방법과 유사하다고 할 수 있다. 그러나 통계적 결합은 서로 다른 원천에서 얻어진 파일로부터 변수 값을 얻는 방법이고, 결합 후 새로 얻어진 자료는 그 크기가 커지기도 하므로 결측치 대입과는 차이가 있다.

하지만 이러한 차이점에도 불구하고 통계적 결합과 결측치 대입 방법의 유사한 특성을 이용하는 통계적 결합에서, 자료의 특성을 잘 파악하여 결측치 대입 방법을 적절히 활용 한다면 자료 결합의 정확성을 높이는 좋은 방안이 될 수 있을 것이다(박상미, 2007).

<table>
<tr><th>종류</th><th>설명</th></tr>
<tr><td><통계적 결합></td><td> 
</td></tr>
<tr><td><결측치 대입></td><td> 
</td></tr>
</table>

<그림 2-2> 결측치 대입과 통계적 결합의 비교

2.3 제약이 없는(unconstrained matching) 결합과 제약이 있는(constrained matching) 결합

통계적 결합은 그 수행 방법에 따라 제약이 있는 결합과 제약이 없는 결합으로 구분된다.

이때 제약이란 제공파일의 개체들이 결합에 사용되는 횟수에 대한 제한을 뜻한다. 제약이 없는 결합은 제공파일의 개체들이 결합에 사용되는 횟수에 상관없이 가장 가까운 개체와 결합한다는 장점이 있지만, 결합 후 결합변수 Y의 분포가 결합 전의 제공파일에서의 분포와 다르다는 단점이 있다. 이러한 단점을 해결하기 위해서 제공파일의 개체가 결합 과정에서 동일한 횟수만큼 사용되도록 하는 제약이 있는 방법을 생각할 수 있다.

이 방법은 제공파일에서의 결합변수의 분포가 결합된 파일에서도 유지된다는 장점이 있지만, 모든 개체가 동일하게 결합해야 하므로 공통변수들 간의 거리가 멀어도 결합이 된다는 단점을 갖는다(Rasseler, 2002).

2.4 통계적 결합의 타당성

통계적 결합의 이점은 모든 변수들에 대한 정보를 포함하는 완전한 데이터를 생성하는 것이다.

결합을 통해 정확한 자료를 얻기 위해서는 통계적 결합 방법의 타당성을 평가할 수 있는 다음의 네 가지 단계를 만족해야 한다.

- 1단계

결합된 값은 참값과 일치해야 한다. 즉 $\tilde{y}_i = y_i, i = 1, 2, \ldots, n_A$ 이어야 한다. 이처럼 결합된 값이 참값과 일치하는 것을 “명중(hit)”이라 하고, 계산을 통해 “명중률(hit rate)”를 구할 수 있다. 결합변수 Y가 p개인 경우

에는 모든 변수가 정확히 일치해야만 명중했다고 할 수 있을 것이다.

- 2단계

결합된 파일에서의 변수의 결합분포는 실제 결합분포를 잘 반영해야 한다. 즉, $\tilde{f}_{X,Y,Z} = f_{X,Y,Z}$ 성립해야 한다.

통계적 결합의 가장 중요한 목적은 $f_{X,Y,Z}$ 분포로부터 얻은 자료와 동일한 자료를 얻는 것이다. 결합파일에서 결합분포를 유지하기 위해서는 공통변수 Z가 주어졌을 때 X와 Y가 조건부 독립(conditional independent)이어야 한다.

$$f_{X,Y|Z} = f_{X|Z} f_{Y|Z} = \tilde{f}_{X,Y|Z} \ , \qquad (2.1)$$

실제로 조건부 독립의 가정이 만족되는 경우는 찾기 힘들지만, 통계적 결합에서는 각 파일에서 존재하지 않는 변수는 얻지 못한 것이 아니라 고려되지 않은 변수이므로 변수들 간의 조건부 독립을 가정할 수 있다.

- 3단계

$\widetilde{cov}(X, Y, Z) = cov(X, Y, Z)$와 같이 변수들의 상관관계 구조와 고차원 적률은 통계적 결합 후 결합 파일에서 잘 유지되어야 한다. 또한 주변분포 역시 $\tilde{f}_{Y,Z} = f_{Y,Z}$ & $\tilde{f}_{X,Z} = f_{X,Z}$ 와 같이 결합파일에서 잘 유지되어야 한다.

때때로 분석자는 상관관계 구조에 의한 변수들의 연관성 등에 관심이 있을 수 있기 때문에 적률과 상관관계 구조가 유지되어야 할 필요가 있다. X, Y 두 변수의 공분산은 다음과 같다.

$$cov(X, Y) = E(cov(X, Y|Z)) + cov(E(Y|Z), E(X|Z)), \qquad (2.2)$$

이때, 조건부 독립 가정에 의해 $E(cov(X, Y|Z)) = 0$이 성립하기 때문에 $\widetilde{cov}(X, Y) = cov(E(X|Z), E(Y|Z))$가 만족되어야 한다.

• 4단계

결합 후, 제공파일의 결합변수의 주변분포와 결합분포는 결합된 파일에서 그대로 유지되어야 한다. 그래서 수용파일에 결합변수 Y가 대체될 때 $\tilde{f}_Y = f_Y$ & $\tilde{f}_{Y,Z} = f_{Y,Z}$ 의 성립이 기대된다.

이러한 네 가지 과정을 통해 결합의 타당성을 고려한다.

제3장 통계적 결합의 방법

통계적 결합 기법에 관한 기존 연구들에서는 거리(distance)와 같은 유사성(similarity) 측도를 이용하여 가장 유사한 개체(nearest neighbor)를 찾거나, 회귀분석과 같은 기법을 적용한 데이터 결합 방법이 제안되었다. 최근에는 정성석 등(2004, 2005)이 회귀분석기법에 k-최근접이웃(nearest neighbor ; k-NN)기법을 적용하여 상대적으로 유사한 개체에 대한 정보손실을 줄이는 방법을 제안하였다.

3.1 최근접이웃(nearest neighbor) 방법

일반적으로 데이터 결합의 원리는 두 파일의 개체 간 유사성을 기준으로 이루어진다. 최근접이웃 방법은 공통변수 Z를 이용하여 수용파일의 한 개체와 제공 파일의 모든 개체들 간의 거리를 계산하여, 그 중 가장 가까운 거리를 갖는 제공 파일의 개체를 선택하여 수용 파일의 해당 개체에 추가시킨다. 이와 같은 과정을 수용 파일의 모든 개체에 대해 수행하여 결합파일을 만든다. 이때 거리 함수로는 유클리디안 거리(Euclidean distance), 마할라

노비스 거리(Mahalanobis distance)를 흔히 사용한다.

유클리디안 거리(Euclidean distance) : $D_{ij} = \sqrt{(X_i - X_j)^2}$, (3.1)

마할라노비스 거리(Mahalanobis distance) :

$$D_{ij} = \sqrt{(X_i - X_j)' \Sigma_{XX}^{-1} (X_i - X_j)} . \quad (3.2)$$

(단, 수용파일의 개체 $i = 1, 2, \ldots, n_R$,

제공파일의 개체 $j = 1, 2, \ldots, n_D$,

Σ_{XX} 는 X변수들의 공분산행렬)

유클리드 거리는 각 변수의 분포가 동일하다고 가정하고, 마할라노비스 거리는 변수들 사이의 상관관계를 고려하므로, 상황에 맞는 거리함수를 선택하여 사용하는 것이 바람직하다.

최근접이웃 방법은 공통변수와 결합변수간의 관계를 무시하고 공통변수만을 이용하여 결합하므로 공통변수의 중요도를 반영하지 못한다는 단점이 있다.

3.2 성향점수(propensity score) 방법

제공파일과 수용파일에 각각 1과 0을 갖는 지시변수를 만들고, 이러한 지시변수와 공통변수로 구해진 성향점수(propensity score)를 이용하여 데이터를 결합하는 방법이 성향점수(propensity score) 방법이다. 이 방법은 공통변수가 많거나 공통 변수들이 서로 다른 분포를 갖는 경우 최근접이웃 방법의 사용에 문제가 있을 때 효과적인 대안으로 이용된다.

Rosenbaum 과 Rubin(1983)은 관측연구에서 확률적 배분을 위해 성향점수(propensity score)방법을 사용할 것을 제안하였다(신영곤, 2006). 임상시험에서의 관측연구는 처리그룹과 대조그룹 간의 확률적 배분이 불가능한 경우에 사용되고, 이는 선택편향(selection bias)의 문제를 발생시키는

등의 문제가 있다.

따라서 선택편향을 제거하고 두 집단 간의 공통변수를 균형화 시켜주기 위하여, 가장 비슷한 특징을 갖고 있는 개체로 짝을 만들거나 비슷한 특성을 갖고 있는 집단으로 층화시킨 후 집단 간의 처리 효과를 추정하는 방법을 고려하게 되는 것이다(이관제, 국세정, 2003).

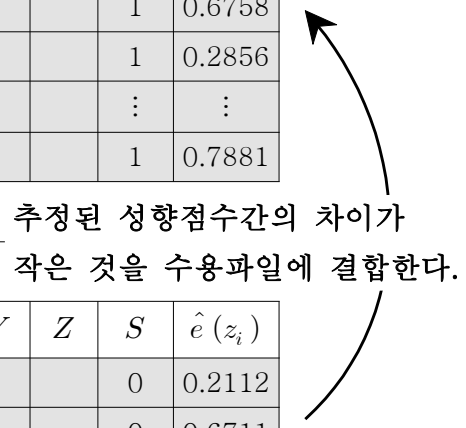
성향점수는 공통변수를 조건으로 실험집단($S=1$) 혹은 대조집단($S=0$)일 조건부 확률을 의미한다. 즉,

$$e(z_i) = P(S=1|Z=z_i) = g(z_i'\beta), \tag{3.3}$$

그리고 식(3.3)과 같이 로짓(logit) 혹은 프로빗(probit) 모형을 이용하여 얻어지는 추정값을 비교하여 결합한다.

$$\hat{e}(z_i) = g(z_i'\hat{\beta}) = \frac{1}{1+e^{-z_i'\hat{\beta}}}, \quad (logit) \qquad \text{또는} \tag{3.4}$$

$$= \Phi(z_i'\hat{\beta}) \quad , \quad (probit) \quad i=1,2,\dots,n \ . \tag{3.5}$$

<table>
<tr><th>종류</th><th>설명</th></tr>
<tr><td>수용파일</td><td>
<table>
<tr><td>i</td><td>X</td><td>Z</td><td>S</td><td>$\hat{e}(z_i)$</td></tr>
<tr><td>1</td><td></td><td></td><td>1</td><td>0.6758</td></tr>
<tr><td>2</td><td></td><td></td><td>1</td><td>0.2856</td></tr>
<tr><td>⋮</td><td></td><td></td><td>⋮</td><td>⋮</td></tr>
<tr><td>n_A</td><td></td><td></td><td>1</td><td>0.7881</td></tr>
</table>


</td></tr>
<tr><td>제공파일</td><td>
<table>
<tr><td>i</td><td>Y</td><td>Z</td><td>S</td><td>$\hat{e}(z_i)$</td></tr>
<tr><td>1</td><td></td><td></td><td>0</td><td>0.2112</td></tr>
<tr><td>2</td><td></td><td></td><td>0</td><td>0.6711</td></tr>
<tr><td>⋮</td><td></td><td></td><td>⋮</td><td>⋮</td></tr>
<tr><td>n_B</td><td></td><td></td><td>0</td><td>0.5502</td></tr>
</table>
</td></tr>
<tr><td>변수설명</td><td>Z : 공통변수
X : 수용파일의 고유변수
Y : 제공파일의 고유변수
S : 지시변수</td></tr>
</table>

<그림 3-1> 성향점수를 이용한 결합방법

3.3 회귀분석(regression)을 이용한 방법

3.1에서 살펴본 바와 같이 최근접이웃 방법은 공통변수와 결합변수간의 연관성을 무시하고 공통변수만을 이용하여 공통변수의 중요도를 반영하지 못하는 단점이 있었다. 이러한 문제점을 개선하기 위해서 Rubin(1986)이 제안한 회귀분석을 이용한 방법을 사용할 수 있다. 이 방법은 결합변수를 종속변수로 하고 공통변수를 설명변수로 하는 선형모형을 통해 예측값을 구하고, 이 예측값으로 구해진 개체간의 거리를 이용하여 결합하는 방법이다.

공통변수를 $Z=\{z_1, z_2, \dots, z_n\}$, 결합변수를 Y라고 두고, 선형 모형식을 다음과 같이 나타낸다.

$$Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \dots + \beta_n z_{n_i} + e_i \,. \tag{3.6}$$

결합변수를 포함하는 제공파일에서 β를 추정하고, 추정된 모수값을 수용파일에 적용하여 회귀식을 구한다. 이렇게 추정된 회귀식을 이용하여 식(3.7)과 같이 예측치간의 근사성을 측정하여 작은 값을 갖는 개체 i와 j를 결합한다.

$$D_{ij} = d(\hat{Y}_i^R - \hat{Y}_j^D) \,. \tag{3.7}$$

여기서, $\hat{Y}_i^R$: 수용파일에서 구해진 예측값

$\hat{Y}_j^D$: 제공파일에서 구해진 예측값

제4장 SAS/AF와 SCL을 통한 시스템 개발

4.1 SAS/AF 및 SCL 소개

SAS/AF(Application Frame)는 그래픽 사용자 인터페이스(graphical user interface)하에서 SAS 소프트웨어에서 제공되는 여러 가지 기능을 이용하여 특별한 목적에 맞게 응용프로그램(application program)을 개발할 수 있도록 지원하는 SAS 제품(product)중 하나이다. 반드시 SAS가 구동되어야만 프로그램이 수행된다는 단점도 있지만, 다른 고급 응용프로그램 언어에 비해서 SAS/AF 응용프로그램이 가지는 가장 큰 장점은 SAS 자체가 가지는 유용한 기능들을 이용하여 강력한 응용프로그램을 비교적 쉽게 개발할 수 있다는 점이다.

SAS/AF 응용프로그램은 주로 엔트리(entry)라고 불리는 구성요소로 이루어지는데, 프레임(frame)과 SCL(Screen Control Language)은 그 중에서도 가장 중요한 엔트리이다.

프레임은 응용프로그램의 흐름에 맞게 개발자가 생성한 화면(window)이라고 할 수 있으며, 사전에 정의된 또는 사용자가 생성한 객체(object)들로 이루어져 있다.

SAS/AF를 이용한 응용프로그램의 개발은 단순히 프레임 엔트리를 작성함으로써 끝나는 것이 아니다. 프레임 엔트리는 실제 응용프로그램 상에서 사용자의 눈에 보이는 겉모양이며, 이 겉모양이 실제로 사용자가 원하는 명령을 수행하도록 하기 위해서는 프레임들을 서로 연결해주고 이들이 작동하도록 명령을 전달하는 SCL(Screen Control Language)이 필요하다. SCL은 프레임과 프레임의 연결, 각 객체의 작동방법(method)등에 대한 프로그램을 저장해 놓은 것으로써, SAS/AF 응용프로그램의 개발은 목적과 흐름에 맞는 프레임과 SCL 엔트리를 작성함으로써 시작된다고 할 수 있다(강현

철 한상태 외, 1999).

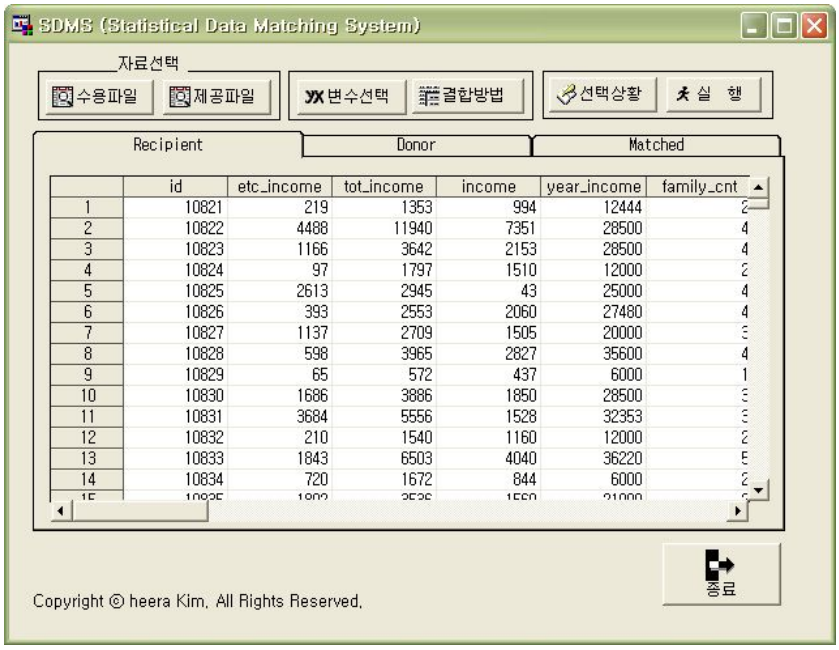
본 연구에서의 결합 시스템은 SAS 9.1을 사용하여 개발되었다.

4.2 시스템 소개

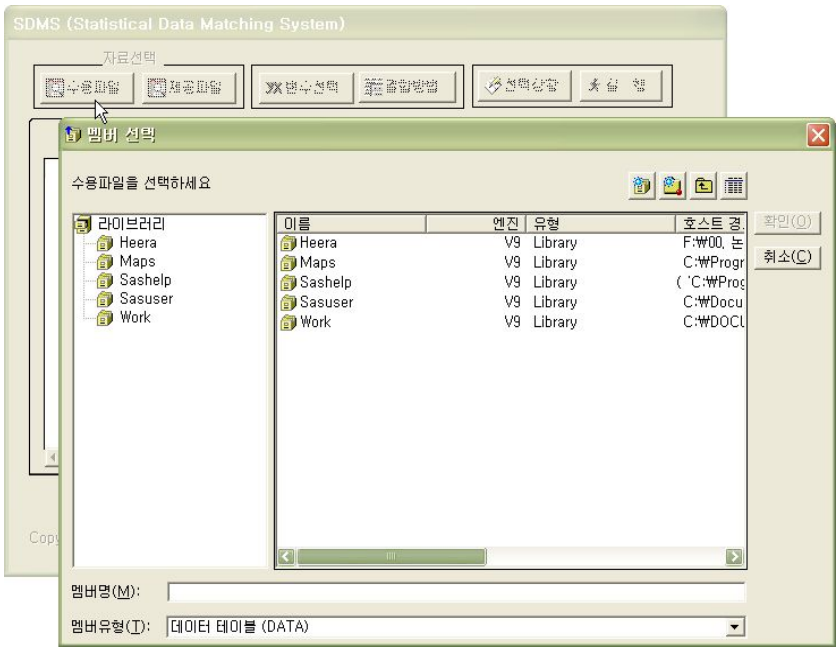
소개할 시스템은 두 개의 파일을 통계적 방법을 이용하여 의미 있게 결합하는 통계적 결합 시스템으로, SDMS(Statistical Data Matching System)라 한다. 이 시스템에서는 결합에 사용될 수용파일과 제공파일을 선택하여 결합변수와 공통변수를 지정하고 최근접이웃, 성향점수, 회귀분석 등의 방법을 선택하여 결합을 실행할 수 있다.

SDMS에서 사용되는 메뉴는 다음과 같다.

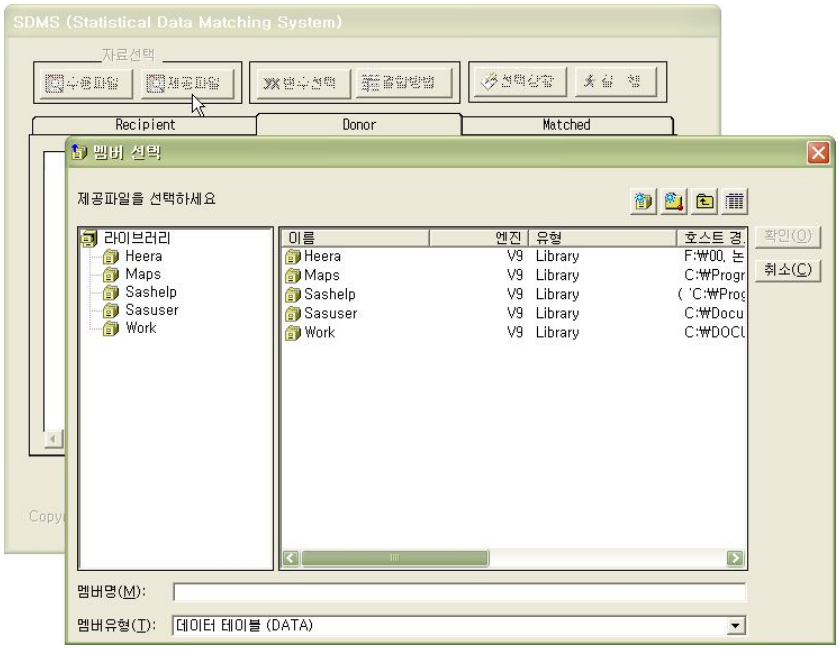
- 수용파일 - 결합변수를 받게 될 수용파일을 선택할 수 있다.
- 제공파일 - 결합변수를 제공할 제공파일을 선택할 수 있다.
- 변수선택 - 선택된 파일로부터 공통변수와 결합변수를 지정할 수 있다.
- 결합방법 - 최근접이웃 방법, 성향점수 방법, 회귀분석 방법 등의 결합 방법을 지정할 수 있다.
- 선택상황 - 선택된 파일명, 공통변수와 결합변수, 결합 방법 등을 확인 할 수 있다.
- 실행 - 선택된 조건들로 통계적 자료결합을 실행한다.
- 데이터 확인탭(tab) - 각 탭(tab)별로 수용파일, 제공파일, 결합파일을 확인할 수 있다.
- 종료 - 시스템을 종료한다.



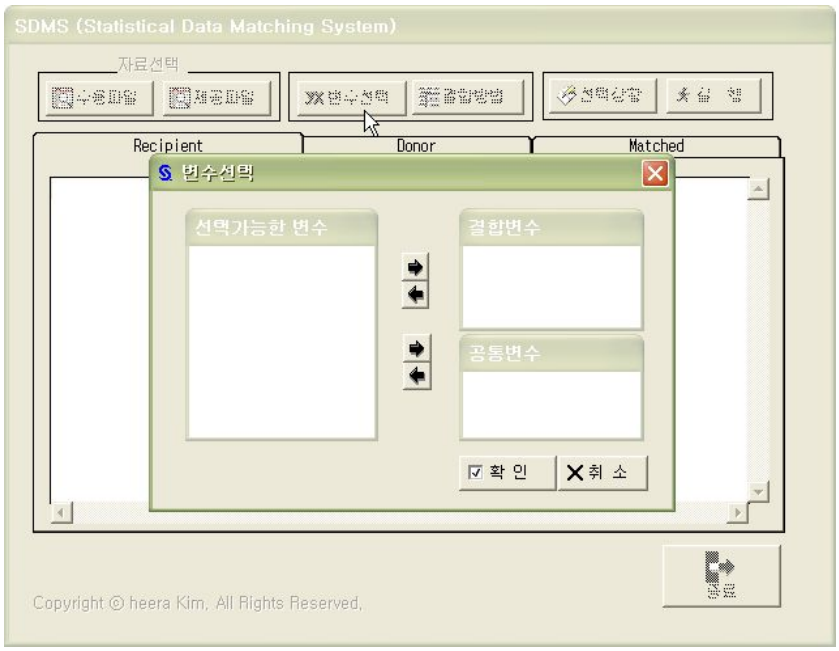
<그림 4-1> 메인화면



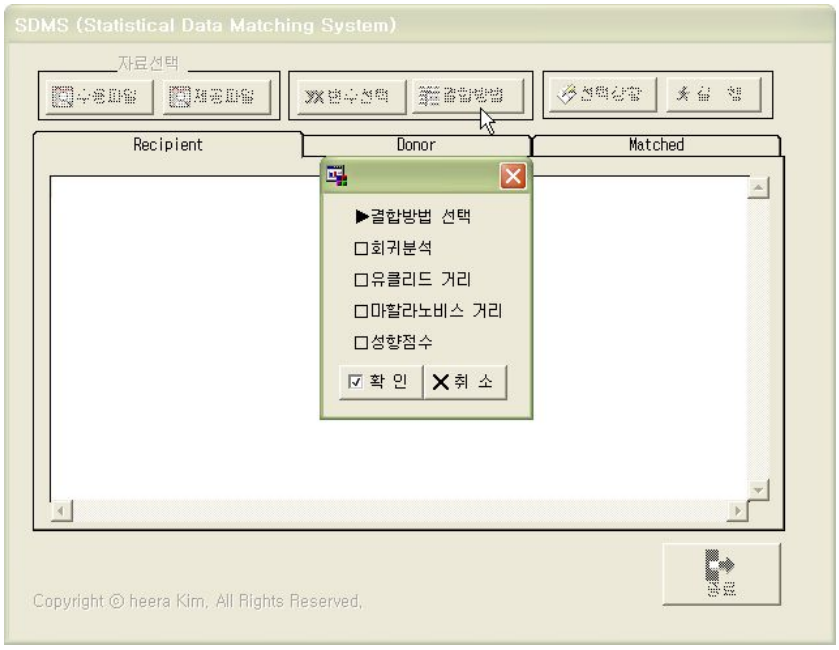
<그림 4-2> 수용파일 선택



<그림 4-3> 제공파일 선택



<그림 4-4> 변수 선택



<그림 4-5> 결합방법 선택

<그림 4-6> 선택상황

4.3 시스템을 이용한 사례분석

4.3.1 자료소개 및 결합절차

시스템을 이용하여 실제 자료를 통한 데이터 결합을 실행해보자. 방법의 정확성을 알아보기 위해 하나의 자료를 임의로 두 개의 자료로 분할하여 결합을 실행 해 보는 폴딩(folding)방법을 이용한다. 두 개의 자료 중 하나의 자료에서는 결합할 변수를 제거한 후 수용파일로 두고 결합을 실행하여 실제 참값과 비교하기로 한다.

데이터는 통계청에서 제공하는 2006년 서울시 도시가계조사 자료를 이용하였다. 자료에는 각 가구당 가구원수, 취업인원수, 가구주에 관한 정보(소득, 연령 등)와 가구의 총 수입, 기타수입, 가계지출 등의 정보가 포함되어 있다.

▸가구수 : 1600가구 (수용파일 800가구, 제공파일 800가구)

▸결합변수 : 기타수입

▸공통변수 : 총수입(소득, 기타수입 등을 포함하는 가계의 모든 수입),
소득(세금 공제전 소득),
연간소득(대략으로 예상한 1년간의 소득),
취업인원수,
가구원수

▸ 사례분석 과정

1단계, 주어진 데이터를 각 800가구를 포함하는 두 개의 파일로 만든다.

2단계, 결합할 변수인 “기타수입” 변수를 제거하여 수용파일을 만든다.

3단계, 나머지 800가구 데이터는 제공파일로 둔다.

4단계, 공통변수를 이용하여 제공파일의 결합변수를 수용파일 에 결합.

5단계, 제공파일로부터 결합된 결합변수와 수용파일에 있던 참값과 비교.

[수용파일]

	공통변수					
obs	취업 인원수	가구원수	총수입	소득	연간소득	기타수입
1	0	2	1353	994	12444	제
⋮	⋮	⋮	⋮	⋮	⋮	거
800	1	4	9662	4760	42600	

[제공파일]

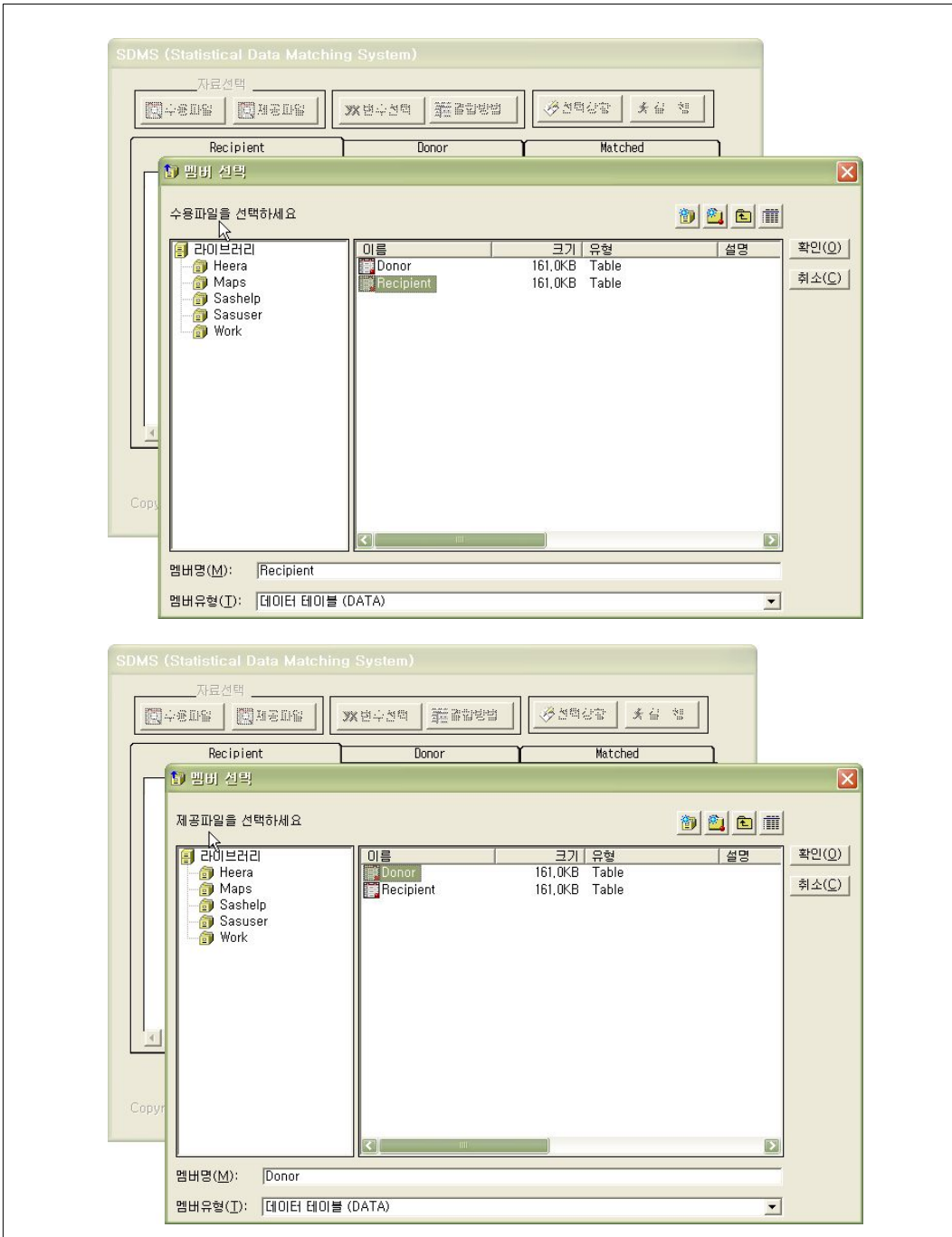
	공통변수					결합변수
obs	취업 인원수	가구원수	총수입	소득	연간소득	기타수입
1	1	3	3808	2100	42000	1248
⋮	⋮	⋮	⋮	⋮	⋮	⋮
800	5	6	6863	6272	85000	284

<그림 4-7> 수용파일과 제공파일

4.3.2 시스템을 이용한 결합절차

▸ 결합에 사용될 데이터 선택 <그림 4-8>

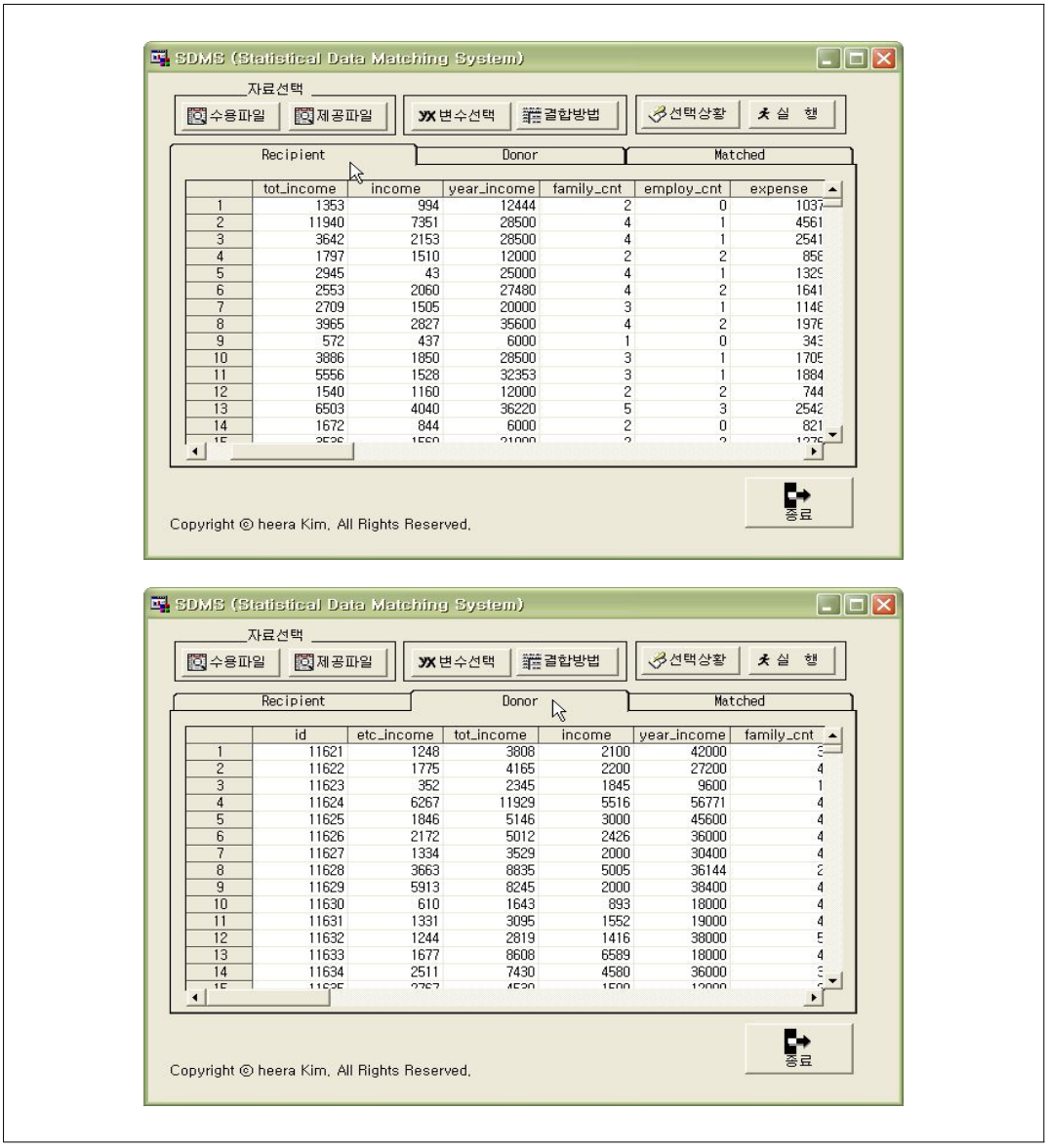
통계적 결합의 첫 단계로써 수용파일과 제공파일로 사용될 데이터를 선택해야 한다. 메인화면의 “수용파일”, “제공파일” 버튼은 SAS에 설정되어 있는 라이브러리상의 데이터를 수용파일과 제공파일로 선택할 수 있다.



<그림 4-8> 수용파일과 제공파일 선택

▸ 선택된 데이터의 확인 <그림 4-9>

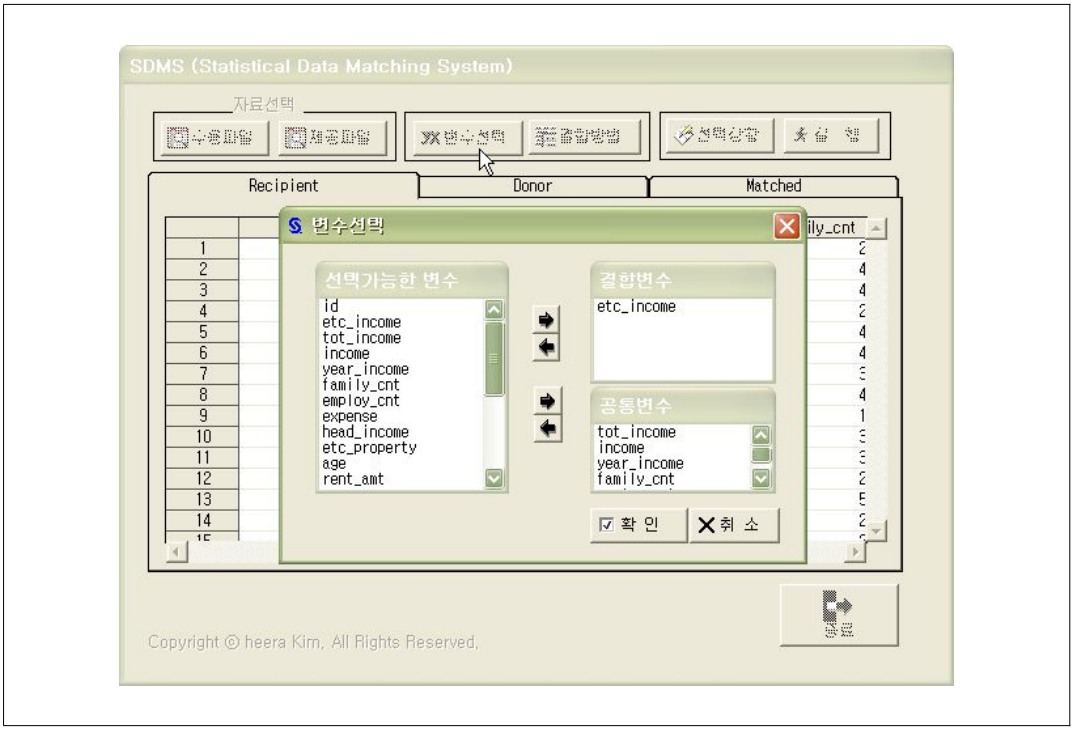
수용파일과 제공파일을 설정한 후 “Recipient”, “Donor” 탭(tab)을 선택 하면 수용파일과 제공파일로 설정된 데이터를 확인할 수 있다.



<그림 4-9> 데이터 확인

▸ 공통변수와 결합변수의 선택 <그림 4-10>

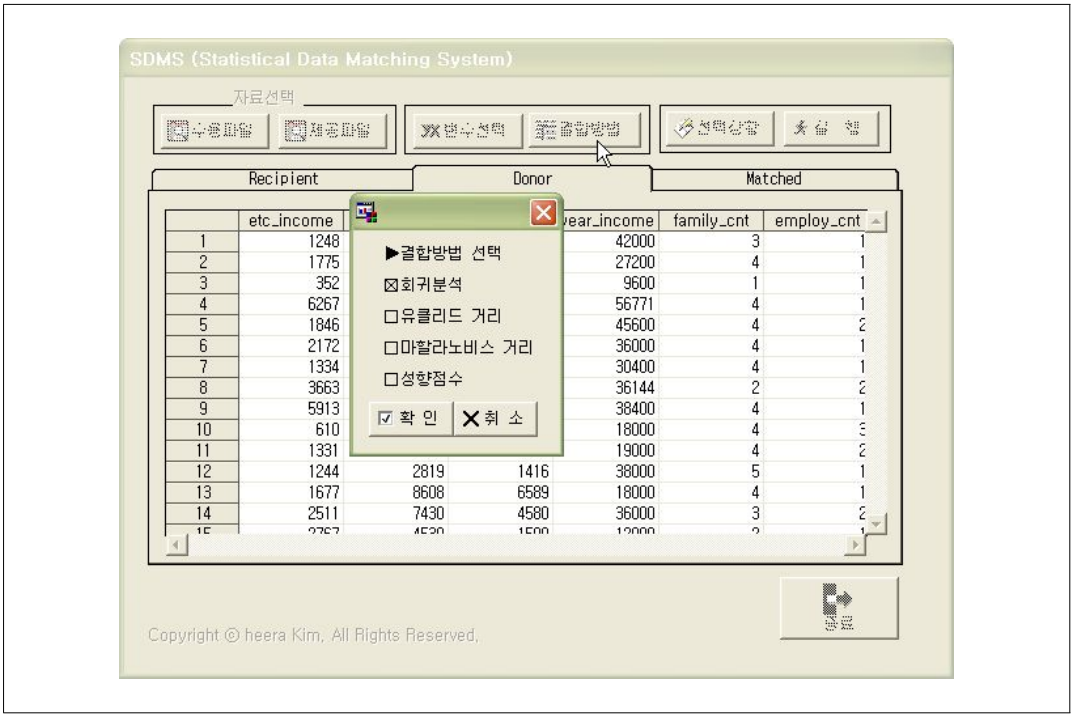
수용파일과 제공파일 설정 후 통계적 자료 결합에 필요한 공통변수와 수용파일에 결합 될 결합변수를 “변수선택” 버튼으로 설정한다.



<그림 4-10> 변수 선택

▸ 결합에 사용할 방법 선택 <그림 4-11>

공통변수과 결합변수의 설정을 마친 후 통계적 결합에 사용하고자 하는 결합방법을 “결합방법” 버튼으로 선택한다.



<그림 4-11> 결합방법 선택

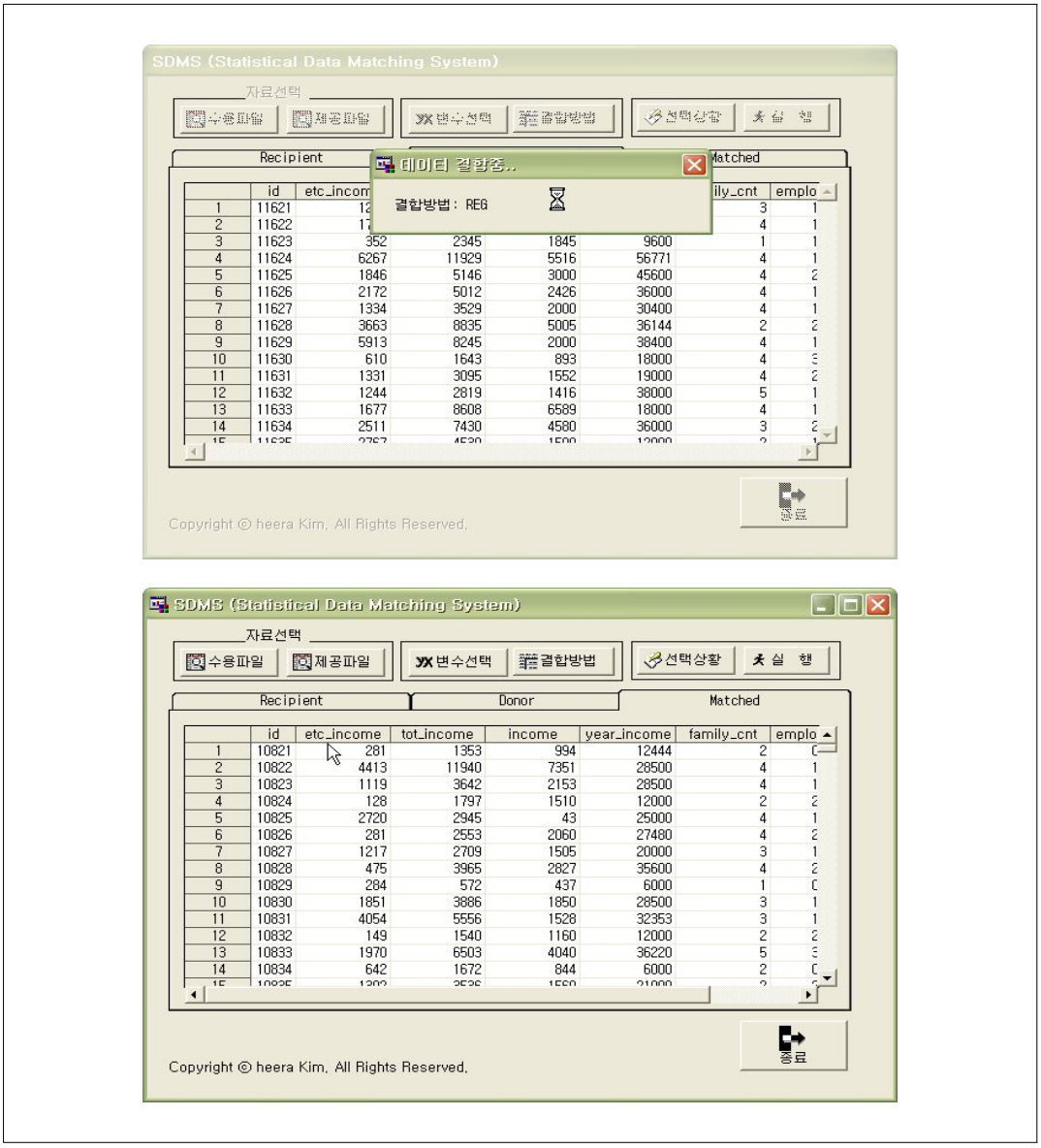
▸ 선택된 사항들의 확인 <그림 4-12>

수용파일과 제공파일 설정, 변수선택, 결합방법 선택을 마친 후 통계적 결합을 실행하기에 앞서, 현재까지의 설정을 “선택상황” 버튼을 통해 확인할 수 있다.

<그림 4-12> 선택상황 확인

▸ <그림 4-13>

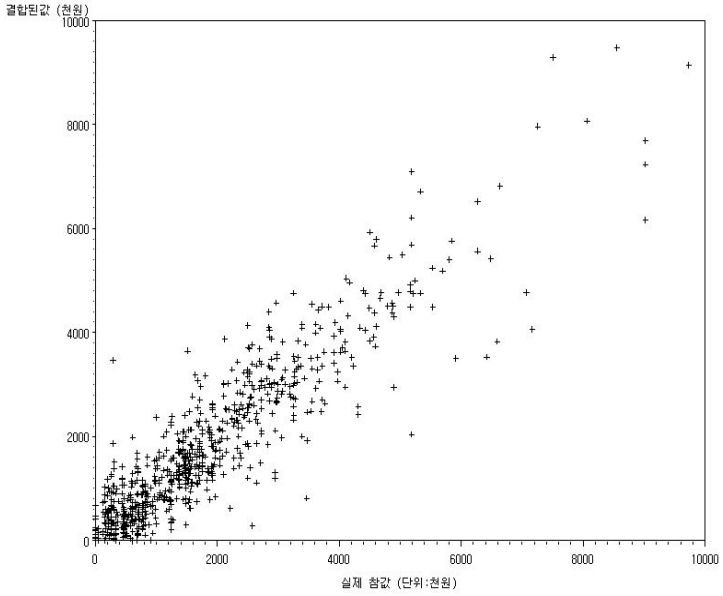
마지막으로 “실행” 버튼을 누르면, 앞서 선택한 조건들을 바탕으로 통계적 결합이 실행되고, “Matched” 탭(tab)에 결합된 파일이 표시된다.



<그림 4-13> 실행 및 결합결과

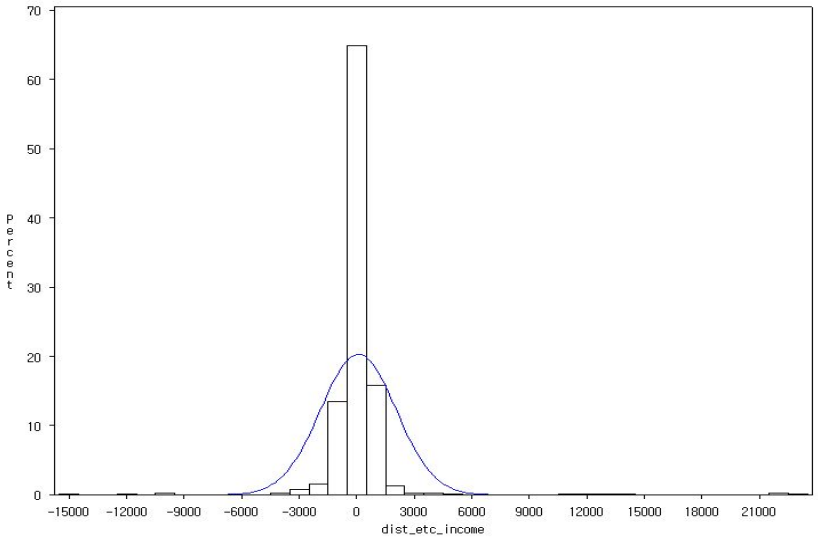
4.4 결합결과 비교

통계적 결합을 수행한 후 결합된 변수의 값과 처음 수용파일에서 인위적으로 제거했던 실제 참값의 분포와, 결합된 값과 참값과의 차이를 그린 그래프를 통해 원래 값에 얼마나 근접하게 결합되었는지 살펴보았다.



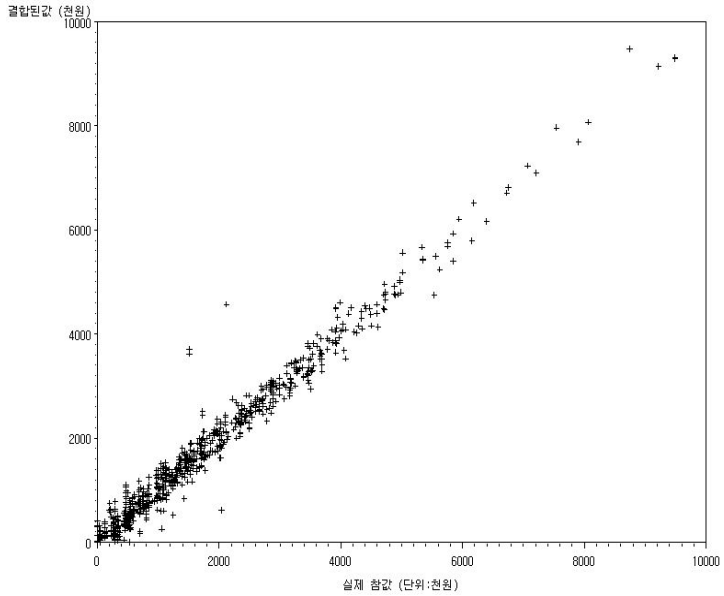
<그림 4-14> 유클리드 거리를 이용한 결합 결과

<그림 4-14>은 유클리드 거리를 사용하여 결합된 파일의 결합된 변수값(기타수입)과 수용파일에서 인위적으로 제거했던 참값의 분포를 그린 것이다. 몇몇 특이값들을 제외하고는 비교적 대칭을 이루고 있는 형태를 확인할 수 있다.



<그림 4-15> 결합된 변수값의 참값과 차이 분포(단위: 천원)

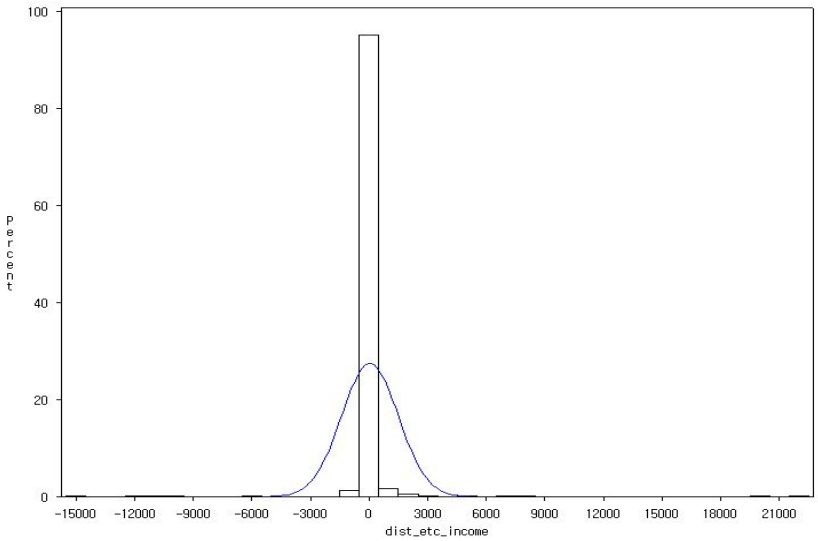
또한 결합된 변수값과 참값과의 차이를 분포로 나타낸 <그림 4-15>은 그 차이가 100만원 내외인 값을 중심으로 하는 분포를 보이고 있다. 참값과의 차이가 100만원으로 조금 크다고 볼 수 있지만 100만원보다 더 큰 차이를 보이는 개체는 거의 없으므로 대체로 결합이 잘 되었다고 볼 수 있다.



<그림 4-16> 회귀분석을 이용한 결합 결과

<그림 4-16>은 회귀분석을 이용한 방법으로 결합된 변수값(기타수입)과 수용파일에서 인위적으로 제거했던 참값의 분포를 그린 것이다.

육안으로도 확실한 대칭을 이루고 있는 분포를 확인할 수 있으며 특이값도 크게 벗어나지 않고 있으므로 결합이 정확결합에 가깝다고 볼 수 있다.



<그림 4-17> 결합된 변수의 참값과의 차이 분포(단위: 천원)

결합된 변수값과 참값과의 차이를 분포로 나타낸 <그림 4-17>은 차이가 100만원 미만인 가까운 개체가 95% 이상을 차지하고 있음을 나타내고 있다. 이러한 결과로부터 회귀분석을 이용한 방법으로 결합된 데이터는 참값과의 차이가 거의 없이 결합이 매우 잘 되었다고 볼 수 있다.

위에서 확인한 결과로 유클리드 거리를 이용한 결합과 회귀분석 방법을 이용한 결합을 비교해본다. 결합된 값과 참값의 분포는 육안으로 확연한 차이를 확인하기 힘들지만, 결합된 변수와 참값과의 차이를 분포로 나타낸 그림으로는 확연한 차이를 확인할 수 있다. 유클리드 거리를 이용한 방법은 값의 차이가 대부분 100만원에 결합이 되어있지만, 회귀분석 방법을 이용한 결합은 그 차이가 100만원 미만인 개체가 대부분을 차지하고 있다. 이러한 방법에 따른 결합결과의 차이는 결합변수로 사용한 “기타수입”이 공통변수로 사용된 “총수입”, “연소득”과의 연관성이 크기 때문에, 변수간의 연관성까지 고려한 회귀분석을 이용한 결합 결과가 더 정확한 것으로 추측된다.

제5장 결론 및 향후 연구방향

본 연구에서는 SAS/AF 와 SCL을 이용하여 통계적 데이터 결합 시스템 SDMS (Statistical Data Matching System)을 개발하였다. 개발된 시스템을 이용하여 통계청에서 제공하는 2007년 서울 도시가계조사 자료를 임의로 두 개의 파일로 나누어 각각 수용파일과 제공파일로 두고, 수용파일에서는 결합하고자 하는 결합변수를 제거한 후 SDMS를 이용하여 통계적 자료결합을 실시하였다.

SDMS에서 제공하는 방법 중 유클리드 거리와 회귀분석을 이용한 방법을 실행하여 실제자료를 결합한 결과, 유클리드 거리를 이용한 방법 보다는 회귀분석을 이용한 방법을 통한 결합이 실제 값과 더 가깝게 결합한 것을 확인하였다. 이러한 결과는 회귀분석을 이용한 방법이 변수간의 연관성까지 고려하였기 때문인 것으로 판단된다.

SDMS는 통계적 결합방법에 대해 전문지식이 없는 사용자들에게는 편리한 시스템일 것이다. 자료 결합에는 결합할 데이터의 특성, 결합변수와 공통변수로 사용되는 변수들의 유형, 결합에 제약을 둘 것인지 등에 따라 다양한 통계적 자료결합 방법이 존재한다. 그러나 현재 구축된 SDMS에는 선택할 수 있는 조건이나 결합방법이 다양하지 않으므로 상세하게 조건을 선택하고 다양한 결합방법을 사용할 수 있도록 하기 위한 연구가 추후 필요할 것이다.

Abstract

A Study on Statistical Matching System using SAS/AF

Hee-ra Kim
Department of Statistics
The Graduate School
Sungshin Women's University

In data analyzing, the availability of data is a very important factor. Generally, analysis data could come from single or various sources, which would be similar or not. In fact, it is rare that a single data set will hold all that we need to answer the question facing. Therefore a key issue is an ability to deal with conflicting data. It is necessary to match the data from various sources for assembling the meaningful information.

As a method of the matching, statistical matching is used for the production of a comprehensive data files from data in various sources. However, it is difficult for non-experts to use statistical data matching techniques because the techniques need statistical knowledge and programing abilities.

This thesis has introduced an overview of statistical matching and developed a statistical data matching system by using SAS/AF and SCL. Finally we have applied the system to the "2006 city family survey of Korea".

참 고 문 헌

[1] 강현철, 한상태, 서혜선, 정형철, 최보승(1999), “*SAS/AF 와 SCL을 이용한 통계적 정보분석시스템 개발*”, 자유아카데미, 서울.

[2] 박상미 (2007), “변수설명력에 따른 데이터 결합의 정확도 비교”, *고려대학교 석사학위논문*.

[3] 신영곤 (2006), “성향점수(propensity score)를 이용한 직업간 혈압차에 대한 비교 연구”, *고려대학교 석사학위논문*.

[4] 이관제, 국세정 (2003), "Propensity Score Matching 방법에 의한 실업자 직업훈련 사업의 효과성 평가", *한국행정학보*, Vol. 7, No. 3, p.181-199.

[5] 정성석, 김순영, 김현진 (2004), “데이터 보강을 위한 데이터 통합기법에 관한 연구”, *응용통계연구*, Vol. 17, No. 3, p.605-617.

[6] Rasseler, S. (2002), “*Statistical Matching : A frequentist theory, practical application and alternative Bayesian approachs*”, Lecture Note in Statistics 168, Springer, New York.

[7] Rosenbaum, P. R., and Rubin, D. B. (1983), “The central role of the propensity score in observational studies for causal effects”. *Biometrika*, Vol. 70, p.41-55.

[8] Rubin, D. B. (1986), “Statistical matching using file concatenation with adjusted weights and multiple imputation”, *Journal of Business & Economic Statistics*, Vol. 4, No. 1, p.87-94.

[9] Saporta, G. (2002), “Data fusion and data grafting”, *Computational Statistics & Data Analysis*, Vol, 38, p.465-473.

<부 록>

1. 메인화면

```
Length dataset1 $20 dataset2 $20 ;
_frame_=_frame_;

INIT:
 Call Send (_frame_, '_GET_WIDGET_', 'table1', tblid1);
 comlst=Makelist();      /* 공통변수 저장 */
 spcflst=Makelist();     /* 결합변수 저장 */
 mat_lat=Makelist();     /* 결합방법 저장 */
RETURN;

MAIN: TERM: RETURN;

TAB:
 Call Notify ('tab', '_GET_ACTIVE_TAB_', tabid);
 select(tabid);
 when(1) do;
  call notify('table1', '_set_dataset_', dataset1);
 end;
 when(2) do;
  call notify('table1', '_set_dataset_', 'dataset2);
 end;
 when(3) do;
  call notify('table1', '_set_dataset_',
'work.matched_file');
 end;
 otherwise;
 end;
RETURN;

VIEW:
 Call Display('SDMS040.frame', dataset1, dataset2,
            comlst, spcflst, mat_lst);
RETURN;

SETDATA1:
 dataset1=Dirlist('', 'data', 1, 'Y', ' ', ' ', '수용파일을
            선택하세요');
RETURN;

SETDATA2:
 dataset2=Dirlist('', 'data', 1, 'Y', ' ', ' ', '수용파일을
            선택하세요');
RETURN;

VAR:
 if not exist(dataset2, 'data') then do;
 _msg_='선택된 데이터가 없습니다.' ;
 return;
 end;
 Call Display('SDMS020.frame', comlst, spcflst,
            dataset2);
RETURN;

METHODS:
 Call Display(SDMS030.frame', mat_lst);
RETURN;

SUBMIT:
 Call Send(tblid1, '_SET_DATASET_', ' ');
 Call Display('SDMS050.frame' , dataset1, dataset2,
            comlst, spcflst, mat_lst);
 tab=1;
 link tab;
RETURN;

END:
 Call Execcmd('cancel');
RETURN;
```

2. 변수선택

```
Entry   comlst 8  spcflst 8  dataset2 $20;

INIT:
 list1=Makelist();
 list2=Makelist();
 list3=Makelist();

 if listlen(spcflst) then list2=copylist(spcflst);
 if listlen(comlst) then list3=copylist(comlst);

 Call Notify ('list1' , '_REPOPULATE_');
 Call Notify ('list2' , '_REPOPULATE_');
 Call Notify ('list3' , '_REPOPULATE_');

 link fill_var;
RETURN;

MAIN:  TERM:  RETURN;

OK:
 spcflst=copylist(list2);
 comlst=copylist(list3);
 Call Execcmd('cancel');
RETURN;

CANCEL:
 Call Execcmd('cancel');
RETURN;
MENU1:
 Call Notify ('menu1' , '_GET_LAST_SEL_' , is ,
                sel, item);
 select(item);
  when('RIGHT1') link right1;
  when('LEFT1') link left1;
  otherwise;
 end;
RETURN;

RIGHT1:
 Call Notify ('list1' , '_GET_NSELECT_', n);
 if n=0 then return;
 do si=1 to n;
  flag=1;
  Call Notify ('list1' , '_SELECTED_', si, ith);
  Call Notify ('list1' , '_GET_TEXT_', ith, item);
  do sj=1 to listlen(list2);
   if GetitemC(list2, sj)=item then flag=0;
  end;
  if flag then
   rc=InsertC(list2, item, -1, Nameitem(list1, ith));
 end;
 Call Notify ('list1' , '_REPOPULATE_');
 Call Notify ('list2' , '_REPOPULATE_');
RETURN;

LEFT1:
 if templist > 0 then rc=clearlist(templist);
 else templist=Makelist();
 Call Notify ('list2' , '_GET_NSELECT_', n);
 do si=1 to n;
  Call Notify ('list2', '_SELECTED_', si, ith);
  rc=InsertN(templist, ith, -1);
 end;
 rc=sortlist(templist, 'descending');
 do si=1 to listlen(templist);
  rc=delitem(list2, GetitemN(templist, si));
 end;
 Call Notify('list1', '_REPOPULATE_');
 Call Notify('list2', '_REPOPULATE_');
RETURN;

MENU2:
 Call Notify ('menu2', '_GET_LAST_SEL_', is,  sel,
                item);
 select(item);
  when('RIGHT2') link right2;
  when('LEFT2') link left2;
  otherwise;
 end;
RETURN;

RIGHT2:
 Call Notify('list1', '_GET_NSELECT_', n);
 if n=0 then return;
 do si=1 to n;
  flag=1;
  Call Notify ('list1', '_SELECTED_', si, ith);
```

```
   Call Notify ('list1', '_GET_TEXT_', ith, item);
   do sj=1 to listlen(list3);
    if GetitemC(list3, sj)=item then flag=0;
   end;
   if flag then
    rc=InsertC(list3, item, -1, Nameitem(list1, ith));
  end;
  Call Notify('list1', '_REPOPULATE_');
  Call Notify('list3', '_REPOPULATE_');
RETURN;

LEFT2:
  if templist >0 then rc=clearlist(templist);
  else templist=Makelist();
  Call Notify ('list3', '_GET_NSELECT_', n);
  do si=1 to n;
   Call Notify ('list3', '_SELECTED_', si, ith);
   rc=InsertN(templist, ith, -1);
  end;
  rc=sortlist(templist, 'descending');
  do si=1 to listlen(templist);
   rc=delitem(list3, GetitemN(templist, si));
  end;
  Call Notify('list1', '_REPOPULATE_');
  Call Notify('list3', '_REPOPULATE_');
RETURN;

FILL_VAR:
  if not exist(dataset2, 'data') then do;
    _msg_='선택한 데이터가 없습니다.' ;
   return;
  end;

  dsid=Open(dataset2, 'I');
  if not dsid then do;
   _msg_='데이터를 열 수 없습니다.';
   return;
  end;

  nvars=Attrn(dsid, 'NVARS');
  do si=1 to nvars;
  if vartype(dsid, si)='N' then do;
   if varlabel(dsid, si)=' ' then
    rc=InsertC(list1, varname(dsid, si), -1,
                varname(dsid, si));
else   rc=InsertC(list1, varname(dsid, si), -1,
                varname(dsid, si));
   end;
  end;
  dsid=close(dsid);
  Call Notify ('list1', '_REPOPULATE_');
RETURN;
```

3. 결합방법 선택

```
Entry  mat_lst 8;

INIT:
  if listlen(mat_lst) > 0 then do;
   do si=1 to listlen(mat_lst);
    if Nameitem(mat_lst, si) ne ' ' then
     Call Notify (Nameitem(mat_lst, si),
                  '_ACTIVATE_', 1);
   end;
  end;
RETURN;

MAIN:  TERM:  RETURN;

OK:
  rc=clearlist(mat_lst);
  if euclid ne ' ' then
   rc=InsertC(mat_lst, 'EUCLID', -1, 'EUCLID' );
  if reg ne ' ' then
   rc=InsertC(mat_lst, 'REG', -1, 'REG' );
  if psm ne ' ' then
   rc=InsertC(mat_lst, 'PSM', -1, 'PSM' );
  if mahala ne ' ' then
   rc=InsertC(mat_lst, 'MAHALA', -1, 'MAHALA' );
  Call Execcmd('end');
RETURN;

CANCEL:
  Call Execcmd('cancel');
RETURN;
```

4. 선택상황 확인

```
Entry dataset1 $20   dataset2 $20   comlst 8   spcflst 8
        mat_lst 8;

INIT:
  dsname=dataset1 & dataset2;
  if listlen(spcflst) > 0 then
    var=GetitemC(spcflst, 1);
  Call Notify ('comlst', '_REPOPULATE_')
  Call Notify ('mat_lst', '_REPOPULATE_')
RETURN;

MAIN:   TERM:   RETURN;

OK:
  Call Execcmd('end');
RETURN;
```

5. 실행

```
Entry  dataset1 $20   dataset2 $20   comlst 8   spclst 8
         mat_lst 8;
Length  mat $20    com $200   spcf $200;
Length  com1 $200   com2 $200   com3 $200
          matvar1 $200;
_frame_ = _frame_;

INIT:
  message='결합을 시작합니다.';
  Call Send(_fram_, '__REFRESH_');
  if exist ('work.M_info', 'DATA') then
     rc=delete('work.M_info', 'DATA');

SUBMIT CONTINUE;
proc datasets kill; run; quit;
%inc 'c:\Program Files\SAS\SAS
        9.1\core\sasmacro\xmacro.sas';
%inc 'c:\Program Files\SAS\SAS
        9.1\core\sasmacro\stdize.sas';
%inc 'c:\Program Files\SAS\SAS
        9.1\core\sasmacro\distance.sas';
ENDSUBMIT;
  if listlen (mat_lst) <= 0 then Call Execcmd('end');
  if listlen (comlst) <= 0 then Call Execcmd('end');
  if listlen (spclst) <= 0 then Call Execcmd('end');

  com=' ';
  do sj=1 to listlen(comlst);
    com=com||' '||Nameitem(comlst, sj);
  end;
  spcf = Nameitem (spcflst, 1);
  do si=1 to listlen(mat_lst);
    mat=Nameitem(mat_lst, si);

  message='결합방법 : ' ||mat ;
    Call Send(_frame_, '_REFRESH_');
    link mat;
  end;
  Call Execcmd('end');
RETURN;

MAIN:   TERM:   RETURN;

MAT:
/*************** 회귀분석 방법 ***************/
  if mat='REG' then do;

SUBMIT CONTINUE;
DATA work.donor;     SET &dataset2;   gb='2';   RUN;
DATA work.recipient;  SET &dataset1;   gb='1';  RUN;
DATA work.raw_data;   SET recipient donor;      RUN;

/* 제공파일로부터 회귀분석 수행 */
PROC REG DATA=donor
    OUTEST=beta(drop=_model_ _type_ _depvar_
                          _rmse_ &spcf) ;
    MODEL &spcf=&com / p noint noprint ;
    ID id ;
    OUTPUT OUT=d_reg_out p=pred;
RUN;
ENDSUBMIT;

/* 제공파일의 회귀분석으로 얻어진 모수를 수용파
일에 매칭 */
com1=' ';
do si=1 to listlen(comlst);
com1=com1||'
```

```
'||Nameitem(comlst,si)||'1'||'='||Nameitem(comlst,si)||';';
end;

SUBMIT CONTINUE;
DATA beta1(drop=&com);
   SET beta;    &com1;    RUN;
ENDSUBMIT;

com2=' ';
do sk=1 to listlen(comlst);
com2=com2||'
'||Nameitem(comlst,sk)||'1'||'='||Nameitem(comlst,sk)||'1;';
end;

SUBMIT CONTINUE;
PROC sql;
   CREATE TABLE recipient_beta AS
   SELECT a.*, b.*
   FROM recipient as a , beta1 as b
;QUIT;
ENDSUBMIT;

com3=' ';
do sl=1 to listlen(comlst);
com3=com3||'+'||Nameitem(comlst,sl)||'1'||'*'||Nameitem(c
omlst,sl) ;
end;

SUBMIT CONTINUE;
DATA r_reg;
   SET recipient_beta;   pred=0 &com3;
RUN;
DATA tot_reg;
   SET r_reg(keep=id pred)    d_reg_out(keep=id pred)  ;
RUN;

/* 거리 매크로 실행 */
%distance (data=tot_reg, id=id, options=nomiss,
out=reg_dist , shape=square, method=euclid, var=pred);

PROC sql;
   CREATE TABLE reg_dist1 AF
   SELET a.*
   FROM reg_dist AS a LEFT JOIN recipient AS b
   ON a.id=b.id
   WHERE b.gb='1'
;QUIT;

PROC TRANSPOSE DATA=reg_dist1 OUT=reg_dist2;
   BY id;
RUN;
DATA reg_dist3(drop=_name_ col1);
   SET reg_dist2;      id2=_NAME_;     value=col1;
RUN;

PROC SQL;
  CREATE TABLE  reg_dist4 AS
   SELECT a.id, a.id2, a.value
   FROM reg_dist3 as a LEFT JOIN donor AS b
   ON a.id2=b.id
   WHERE b.gb='2'
   GROUP by 1
   HAVING min(value)=value
   ORDER by a.id, a.id2
 ;QUIT;

DATA reg_dist5;
   SET reg_dist4;
   BY id;      IF first.id ;
RUN;
ENDSUBMIT;

matvar1=' ';
do sm=1 to listlen(comlst);
matvar1=matvar1||',b.'||Nameitem(comlst,sm) ;
end;

SUBMIT CONTINUE;
PROC SQL;
   CREATE TABLE matched_file AS
   SELECT a.id , c.&spcf   &matvar1
   FROM reg_dist5 as a LEFT JOIN recipient AS b
   ON a.id=b.id
   LEFT JOIN donor AS c   ON a.id2=c.id
   ORDER BY a.id
;QUIT;
```

```
DATA matched_file;
  RETAIN id &spcf &com;  SET matched_file;
RUN;

PROC DATASETS;
  DELETE beta beta1 beta2 d_reg_out raw_data
         recipient_beta reg_dist reg_dist1
         reg_dist2  reg_dist3 reg_dist4 reg_dist5
         r_reg tot_reg _id _idby _name
         _name2 _nomiss _tran_ _var donor
        recipient;
RUN;
ENDSUBMIT;
end;


/************ 유클리드 거리 방법 ************/
else if mat='EUCLID' then do;

SUBMIT CONTINUE;
DATA work.donor;
  SET &dataset2;  gb='2';
RUN;
DATA work.recipient;
  SET &dataset1;  gb='1';
RUN;
DATA work.raw_data;
  SET recipient donor;
RUN;

%distance(data=raw_data, id=id, options=nomiss,
          out=dist, shape=square,  method=euclid,
          var=&com);

PROC SQL;
  CREATE TABLE dist1 AS
  SELECT a.*
  FROM dist as a LEFT JOIN recipient AS b
  ON a.id=b.id
  WHERE b.gb='1'
;QUIT;

PROC TRANSPOSE DATA=dist1 OUT=dist2;
  BY id;
RUN;
DATA dist3(drop=_name_ col1);
  SET dist2;  id2=_NAME_;  value=col1;
RUN;

PROC SQL;
  CREATE TABLE dist4 AS
  SELECT a.id, a.id2, a.value
  FROM dist3 as a LEFT JOIN donor AS b
  ON a.id2=b.id
  WHERE b.gb='2'
  GROUP by 1
  HAVING min(value)=value
  ORDER by a.id, a.id2
;QUIT;

DATA dist5;
  SET dist4;
  BY id;  IF first.id ;
RUN;
ENDSUBMIT;

matvar1=' ';
do sm=1 to listlen(comlst);
matvar1=matvar1||',b.'||Nameitem(comlst,sm) ;
end;

SUBMIT CONTINUE;
PROC SQL;
  CREATE TABLE atched_file AS
  SELECT a.id , c.&spcf  &matvar1
  FROM dist5 as a LEFT JOIN recipient AS b
  ON a.id=b.id
  LEFT JOIN donor AS c  ON a.id2=c.id
  ORDER a.id
;QUIT;

DATA matched_file;
  RETAIN id &spcf &com;  SET matched_file;
RUN;
```

```
PROC DATASETS;
  DELETE dist dist1 raw_data _idby dist2 dist3 dist4
         dist5 _id _name _name2  _nomiss _tran_
         _var donor recipient;
RUN;
ENDSUBMIT;
end;


/*************** 성향점수 방법 ***************/
if mat='PSM' then do;
SUBMIT CONTINUE;
DATA recipient;   SET &dataset1;   gb=1;   RUN;
DATA donor;       SET &dataset2;   gb=0;   RUN;
DATA raw_data;    SET recipient donor;     RUN;

/* 성향점수 계산 */
PROC LOGISTIC DATA=raw_data descending  noprint ;
  MODEL gb=&com
  OUTPUT out=preds pred=propen   ;
RUN;

/* 성향점수의 로짓 */
DATA preds;
  SET preds;
  logit=log(propen/(1-propen));
RUN;

/* 수용파일과 제공파일 분리 */
DATA _recipient;
  SET preds;    IF gb=1;
RUN;
DATA _donor;
SET preds;    IF gb=0;
RUN;

/* matching을 위한 거리계산( D=| P1 - P0 |) */
PROC SQL;
  CREATE TABLE dist AS
  SELECT a.id as r_id , a.logit as r_logit,  b.id as d_id,
         b.logit as d_logit,
         abs(a.logit-b.logit) as dist_logit
  FROM _recipient as a , _donor as b
  GROUP by a.id
    HAVING min(dist_logit)=dist_logit
    ORDER by a.id, b.id
;QUIT;

DATA dist1;
  SET dist;
  IF first.r_id;   BY r_id;
RUN;
ENDSUBMIT;

/* matched file */
matvar1=' ';
do sm=1 to listlen(comlst);
matvar1=matvar1||',b.'||Nameitem(comlst,sm);
end;

SUBMIT CONTINUE;
PROC SQL;
  CREATE TABLE matched_file AS
  SELECT a.r_id as id, c.&spcf &matvar1
  FROM dist1 AS a LEFT JOIN recipient AS b
  ON a.r_id=b.id
  LEFT JOIN donor    AS c ON a.d_id=c.id
  ORDER by 1
;QUIT;

DATA matched_file;
  RETAIN id &spcf &com;    SET matched_file;
RUN;

PROC DATASETS;
  DELETE dist dist1 preds raw_data _donor _recipient
         donor recipient;
RUN;
ENDSUBMIT;
end;
```

감사의 글

책상 위 한가득 쌓인 자료들과 미뤄두었던 수많은 약속들을 남기고 드디어 논문이 완성 되었습니다. 과연 해낼 수 있을지 한치 앞도 내다볼 수 없던 시간들 이었지만, 자신과의 싸움에서 승리하며 손수 만들어낸 이 결과물이 참으로 뿌듯하고 기쁩니다.

부족한 제가 이렇게 논문을 완성할 수 있도록 많은 가르침을 주시고 언제나 든든한 힘이 되어주신 이성건 지도교수님께 깊이 감사드립니다. 활짝 웃으시는 모습만으로도 따뜻함을 주시는 송일성 교수님, 더 넓은 곳으로 나가 많은 것을 경험하라고 좋은 말씀 해 주시는 이해용 교수님, 부족함 많은 제 논문을 꼼꼼히 다듬어주시고 심사 해 주신 이우선 교수님, 또한 정신적인 지주 같은 존재이신 이종협 교수님께 말로 다 할 수 없는 깊은 감사드립니다.

일과 대학원생활을 병행하며 시간에 쫓기고 몸도 힘들었습니다. 이러한 제 상황을 이해해주시고 많은 편의를 봐 주신 삼성테스코 고객가치창조팀 이덕기 이사님과 그 외 식구들에게 감사의 마음을 전합니다. 특히 엄마처럼 때로는 선배처럼 돌봐주신 백승희 대리님, 든든한 큰언니 현정 대리님, 자상한 막내언니 명선언니, 또한 옆에서 많은 힘이 되어준 현아와 미화에게 감사의 뜻을 전합니다.

학부 때부터 대학원까지 함께하며 힘이 되어준 영은, 주현, 바쁜 회사생활에도 후배에게 신경 써 준 가영언니, 논문준비로 힘들어 할 때마다 작은 것부터 세심하게 챙겨준 소중한 후배들 인경, 희원, 논문 쓴다고 매번 약속도 미루며 신경 써주지 못했지만 마음으로 격려해 준 소중한 친구들 민경, 혜연, 승은, 지영, 나의 카운슬러 태경, 대학원 동기가 없어 외로울 때 동기처럼 힘이 되준 진욱오빠, 만호오빠 에게도 감사의 마음 전합니다.

언제나 내 편인 너무 사랑하는 아빠 엄마, 멀리 캐나다에서도 논문 쓰는 언니걱정 많이 해 준 우리 은나에게 감사와 사랑의 마음을 전합니다.

그 외에도 도움 주신 많은 분들께 감사드립니다.